entropy

MDPI

Article

A Novel Approach to Canonical Divergences within Information Geometry

Nihat Ay [1,2,3,*] and Shun-ichi Amari [4]

Received: 12 October 2015 ; Accepted: 25 November 2015 ; Published: 9 December 2015
Academic Editor: Kevin H. Knuth

1 Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, Leipzig 04103 , Germany
2 Faculty of Mathematics and Computer Science, University of Leipzig, PF 100920, Leipzig 04009, Germany
3 Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
4 Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute, Wako-shi Hirosawa 2-1, Saitama 351-0198, Japan; amari@brain.riken.jp
* Correspondence: nay@mis.mpg.de; Tel.: +49-341-9959-547; Fax: +49-341-9959-555

Abstract: A divergence function on a manifold M defines a Riemannian metric g and dually coupled affine connections ∇ and ∇^* on M. When M is dually flat, that is flat with respect to ∇ and ∇^*, a canonical divergence is known, which is uniquely determined from (M, g, ∇, ∇^*). We propose a natural definition of a canonical divergence for a general, not necessarily flat, M by using the geodesic integration of the inverse exponential map. The new definition of a canonical divergence reduces to the known canonical divergence in the case of dual flatness. Finally, we show that the integrability of the inverse exponential map implies the geodesic projection property.

Keywords: information geometry; canonical divergence; relative entropy; α-divergence; α-geodesic; duality; geodesic projection

1. Introduction: Divergence and Dual Geometry

A divergence function $D(p \,\|\, q)$ is a differentiable real-valued function of two points p and q in a manifold M. It satisfies the non-negativity condition

$$D(p \,\|\, q) \geq 0 \tag{1}$$

with equality if and only if $p = q$. Thus, it is a distance-like function, but does not necessarily share all properties of a distance. For instance, it can be asymmetric in p and q. When a coordinate system $\boldsymbol{\xi} : p \mapsto \boldsymbol{\xi}_p = (\xi^1_p, \ldots, \xi^n_p) \in \mathbb{R}^n$ is given in M, we pose one condition that, for two nearby points $\boldsymbol{\xi}_p$ and $\boldsymbol{\xi}_q = \boldsymbol{\xi}_p + \Delta\boldsymbol{\xi}$, D is expanded as

$$D(p \,\|\, q) \;=\; \frac{1}{2} \overset{D}{g}_{ij}(p)\, \Delta\xi^i \Delta\xi^j + O\left(\|\Delta\boldsymbol{\xi}\|^3\right) \tag{2}$$

and $(\overset{D}{g}_{ij}(p))_{ij}$ is a positive definite matrix. Here, the Einstein summation convention is used, which means that summation is taken with respect to any index that appears twice in a term, as a lower as well as an upper index. Throughout the paper, we apply this convention or explicitly use the summation sign. The coefficients $\overset{D}{g}_{ij}$ in Equation (2) define a Riemannian metric $\overset{D}{g}$. Furthermore, the divergence function D allows us to define also a pair of dual affine connections [1]. In order to be

more explicit, we consider coordinates $\xi_p = (\xi_p^1, \ldots, \xi_p^n)$ of p and coordinates $\xi_q = (\xi_q^1, \ldots, \xi_q^n)$ of q and introduce the following simplified notations of differentiation

$$\partial_i = \frac{\partial}{\partial \xi_p^i}, \quad \partial_i' = \frac{\partial}{\partial \xi_q^i} \tag{3}$$

With $D(\xi_p \,\|\, \xi_q) = D(p \,\|\, q)$, the coefficients of the Riemannian metric can be written as

$$\overset{D}{g}_{ij}(p) \;=\; -\partial_i \partial_j' D(\xi_p \,\|\, \xi_q)\Big|_{q=p} \;=\; \partial_i' \partial_j' D(\xi_p \,\|\, \xi_q)\Big|_{q=p} \tag{4}$$

Furthermore, the coefficients

$$\overset{D}{\Gamma}_{ijk}(p) \;=\; -\partial_i \partial_j \partial_k' D(\xi_p \,\|\, \xi_q)\big|_{q=p} \tag{5}$$

$$\overset{D}{\Gamma^*}_{ijk}(p) \;=\; -\partial_i' \partial_j' \partial_k D(\xi_p \,\|\, \xi_q)\Big|_{q=p} \tag{6}$$

define a pair of dual affine connections $\overset{D}{\nabla}$ and $\overset{D}{\nabla^*}$ [1]. The duality of the connections holds with respect to the Riemannian metric $\overset{D}{g}$ in terms of the following condition:

$$X\langle Y, Z\rangle \;=\; \left\langle \overset{D}{\nabla}_X Y, Z \right\rangle + \left\langle Y, \overset{D}{\nabla^*}_X Z \right\rangle \tag{7}$$

for all vector fields X, Y and Z, where the brackets $\langle \cdot, \cdot \rangle$ denote the inner product with respect to $\overset{D}{g}$ [2].

The inverse problem is to find a divergence D which generates a given geometrical structure (M, g, ∇, ∇^*). Matumoto [3] showed that a divergence exists for any such manifold. However, it is not unique and there are infinitely many divergences that give the same geometrical structure. When a manifold is dually flat, a canonical divergence was introduced by Amari and Nagaoka [2], which is a Bregman divergence. Extensions of the canonical divergence within conformal geometry have been studied by Kurose [4] and Matsuzoe [5]. The canonical divergence has nice properties such as the generalized Pythagorean theorem and the geodesic projection theorem. It is an important problem to define a canonical divergence in the general case. The present paper gives an answer to this problem by using the inverse exponential map. We already used the inverse exponential map in our previous work [6], where we studied a different divergence function. We could show that it recovers the metric g in the sense of Equation (4) and has some consistency with the dual connections ∇ and ∇^*. However, it turns out that it does not reduce to the well-established canonical divergence in the dually flat case. The divergence introduced in the present article not only recovers the original geometry directly in terms of Equations (4)–(6), it also coincides with the original canonical divergence in the dually flat case.

2. A New Approach to the General Inverse Problem

We begin with a motivation in terms of a simple example where the manifold is $\mathbb{R}^n$ equipped with the standard Euclidean metric and connection (here, the Levi-Civita connection): Let p be a fixed point in $\mathbb{R}^n$, and consider the vector field pointing to p, that is

$$\mathbb{R}^n \;\to\; \mathbb{R}^n, \quad q \;\mapsto\; p - q \tag{8}$$

Obviously, the vector field Equation (8) can be seen as the negative gradient of the squared distance

$$D_p : \mathbb{R}^n \;\to\; \mathbb{R}, \qquad q \;\mapsto\; D_p(q) \;:=\; D(p \,\|\, q) \;:=\; \frac{1}{2}\|p - q\|^2 \;=\; \frac{1}{2}\sum_{i=1}^{n}(p_i - q_i)^2$$

as potential function, that is

$$p - q \;=\; -\mathrm{grad}_q\, D_p \tag{9}$$

Here, the gradient grad_q is taken with respect to the canonical inner product on $\mathbb{R}^n$.

We shall now generalize the relation Equation (9) between the squared distance D_p and the difference of two points p and q to the more general setting of a differentiable manifold M. Given a fixed point $p \in M$, we want to define a vector field $q \mapsto X(q,p)$, at least in a neighbourhood of p, that corresponds to the difference vector field Equation (8). Obviously, the problem is that the difference $p - q$ is not naturally defined for a general manifold M. We need an affine connection ∇ in order to have a notion of a difference. Given such a connection ∇, for each point $q \in M$ and each direction $X \in T_qM$ we consider the geodesic $\gamma_{q,X}(t)$, with the initial point q and the initial velocity X, that is $\gamma_{q,X}(0) = q$ and $\dot\gamma_{q,X}(0) = X$. If $\gamma_{q,X}(t)$ is defined for all $0 \le t \le 1$, the endpoint $p = \gamma_{q,X}(1)$ is interpreted as the result of a translation of the point q along a straight line in the direction of the vector X. This straightness is expressed in terms of the local coordinates $\boldsymbol{\xi}(t) := (\xi^1(t), \cdots, \xi^n(t)) := \boldsymbol{\xi}(\gamma_{q,X}(t))$ of the geodesic $\gamma_{q,X}$ by the following set of differential equations:

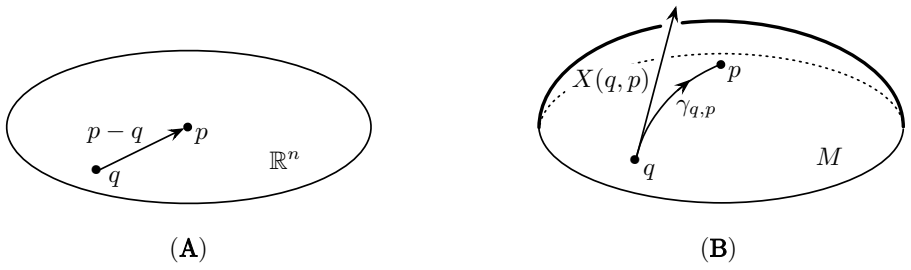
$$\ddot\xi^i(t) + \Gamma^i_{jk}(\boldsymbol{\xi}(t))\,\dot\xi^j(t)\dot\xi^k(t) \;=\; 0\,, \qquad i = 1, \ldots, n \tag{10}$$

The translation of points along geodesics defines a map, the so-called *exponential map*:

$$\exp_q :\; U_q \;\to\; M\,, \quad X \;\mapsto\; \gamma_{q,X}(1) \tag{11}$$

where $U_q \subseteq T_qM$ denotes the set of tangent vectors X, for which the domain of $\gamma_{q,X}$ contains the unit interval $[0,1]$.

Given two points p and q, one can interpret any X with $\exp_q(X) = p$ as a difference vector X that translates q to p. Throughout this paper we assume the existence and uniqueness of such a difference vector, denoted by $X(q,p)$ (see Figure 1).



(**A**) (**B**)

Figure 1. Illustration of (**A**) the difference vector $p - q$ in $\mathbb{R}^n$ pointing from q to p; and (**B**) the difference vector $X(q,p) = \dot\gamma_{q,p}(0)$ as the inverse of the exponential map in q.

This is a strong assumption, which is, however, always locally satisfied. On one hand, we are mainly interested in local properties. On the other hand, although being quite restrictive in general, this property will be satisfied in our information-geometric context, where g is given by the Fisher metric and ∇ is given by the m- and e-connections and their convex combinations, the α-connections.

If we attach to each point $q \in M$ the difference vector $X(q,p)$, we obtain a vector field that corresponds to the vector field Equation (8) in $\mathbb{R}^n$. In order to interpret this vector field as a negative gradient field of a (squared) distance function, and thereby generalize Equation (9), we need a Riemannian metric g on M. Given such a metric, we assume *integrability* of X and ∇, respectively, in the sense that for all p there exists a function D_p satisfying

$$X(q,p) \;=\; -\mathrm{grad}_q\, D_p \tag{12}$$

Here, the Riemannian gradient is taken with respect to g, which is defined by the property that the total differential $d_q D_p$ can be expressed as an inner product:

$$\langle \operatorname{grad}_q D_p, Y \rangle \;=\; d_q D_p(Y)\,, \qquad Y \in T_q M$$

Obviously, if there are functions D_p satisfying the condition of Equation (12) then they are unique up to a constant that can vary with p, and we can therefore assume $D_p(p) = 0$. Throughout the paper we will also use the standard notation $D(p \,\|\, q) = D_p(q)$ of a divergence as a function D of two arguments. In order to recover D from Equation (12) we consider any curve $\gamma : [0,1] \to M$ that connects q with p, that is $\gamma(0) = q$ and $\gamma(1) = p$. We compose the inner product of the curve velocity $\dot\gamma(t)$ with the inverse of the exponential map $X(\gamma(t), p)$ in $\gamma(t)$ and integrate this along the curve:

$$\begin{aligned}
\int_0^1 \langle X(\gamma(t), p), \dot\gamma(t) \rangle \, dt \;&=\; -\int_0^1 \left\langle \operatorname{grad}_{\gamma(t)} D_p, \dot\gamma(t) \right\rangle dt \\
&=\; -\int_0^1 (d_{\gamma(t)} D_p)(\dot\gamma(t))\, dt \\
&=\; -\int_0^1 \frac{d\, D_p \circ \gamma}{d\, t}(t)\, dt \\
&=\; D_p(\gamma(0)) - D_p(\gamma(1)) \\
&=\; D_p(q) - D_p(p) \;=\; D_p(q) \;=\; D(p \,\|\, q)
\end{aligned} \tag{13}$$

In particular, we can apply this derivation to the geodesic connecting q and p even when the integrability of X is not guaranteed and obtain the definition of a general canonical divergence, discussed in more detail in Section 5. Before we treat the general definition of a canonical divergence, however, we discuss important special cases of divergences within the cone of positive measures and the simplex of probability measures included in it. In particular, we verify that the well-known relative entropy (KL-divergence) and the α-entropy (α-divergence) can be derived in terms of Equation (13).

3. Natural Connections for Positive and Probability Measures

3.1. The Fisher Metric and Its Gradients

We represent measures on the set $\{1, \dots, n\}$ as elements of $\mathbb{R}^n$. In this representation, the Dirac measures δ_i, $i = 1, \dots, n$, form the canonical basis of $\mathbb{R}^n$. We consider the n-dimensional cone of positive measures on the set $\{1, \dots, n\}$, defined by

$$M_n \;:=\; \mathbb{R}^n_+ \;=\; \left\{ p = \sum_{i=1}^n p_i\, \delta_i \in \mathbb{R}^n \;:\; p_i > 0 \text{ for all } i \right\}$$

and the corresponding $(n-1)$-dimensional simplex of normalized measures (probability measures) $S_{n-1} \subset M_n$:

$$S_{n-1} \;:=\; \left\{ p = \sum_{i=1}^n p_i\, \delta_i \in \mathbb{R}^n \;:\; p_i > 0 \text{ for all } i, \text{ and } \textstyle\sum_{i=1}^n p_i = 1 \right\}$$

There is a natural Riemannian metric on M_n, called the Fisher metric:

$$g_p(X, Y) \;:=\; \sum_{i=1}^n \frac{1}{p_i}\, X_i\, Y_i\,, \qquad X, Y \in T_p M_n$$

In theoretical biology, the Fisher metric is also known as Shahshahani metric (see [7], Equation (7.48)). Given a point $p \in S_{n-1}$ and a vector $X \in T_p M_n$, its projection onto $T_p S_{n-1}$ with respect to g_p is given by

$$\Pi_p^{\top} X \;=\; \sum_{i=1}^{n} \left(X_i - p_i \sum_{j=1}^{n} X_j \right) \delta_i \tag{14}$$

and the corresponding projection onto the orthogonal complement of $T_p S_{n-1}$ is given by

$$\Pi_p^{\perp} X \;=\; \sum_{i=1}^{n} \left(p_i \sum_{j=1}^{n} X_j \right) \delta_i \tag{15}$$

For a function $V : M_n \to \mathbb{R}$, this metric implies the Riemannian gradient

$$\operatorname{grad}_p V \;=\; \sum_{i=1}^{n} \left(p_i \frac{\partial V}{\partial p_i}(p) \right) \delta_i \tag{16}$$

A vector field

$$X_p \;=\; \sum_{i=1}^{n} p_i \, f_i(p) \, \delta_i \,, \qquad p \in M_n \tag{17}$$

is the gradient of a function V if and only if it satisfies for all i, j

$$\frac{\partial f_i}{\partial p_j} \;=\; \frac{\partial f_j}{\partial p_i} \tag{18}$$

If we consider a function that is defined on S_{n-1}, for instance the restriction of V: $M_n \to \mathbb{R}$ to S_{n-1}, then the vector Equation (16), evaluated in $p \in S_{n-1}$, will not necessarily be an element of $T_p S_{n-1}$. Therefore, in order to evaluate the gradient on S_{n-1}, we have to project the vector Equation (16) onto $T_p S_{n-1}$ with respect to the metric g by using Equation (14). This leads to the following gradient formula for functions on S_{n-1}:

$$\operatorname{grad}_p V \;=\; \sum_{i=1}^{n} p_i \left(\frac{\partial V}{\partial p_i}(p) - \sum_{j=1}^{n} p_j \frac{\partial V}{\partial p_j}(p) \right) \delta_i \,, \qquad p \in S_{n-1} \tag{19}$$

This gives rise to consider general vector fields of the form

$$X_p \;=\; \sum_{i=1}^{n} p_i \left(f_i(p) - \sum_{j=1}^{n} p_j \, f_j(p) \right) \delta_i \,, \qquad p \in S_{n-1} \tag{20}$$

Such a vector field is integrable, in the sense that it is the gradient Equation (19) of a potential function V, if and only if the following condition holds for all i, j, k (see [7], Equation (19.23)):

$$\frac{\partial f_i}{\partial p_j} + \frac{\partial f_j}{\partial p_k} + \frac{\partial f_k}{\partial p_i} \;=\; \frac{\partial f_i}{\partial p_k} + \frac{\partial f_k}{\partial p_j} + \frac{\partial f_j}{\partial p_i} \tag{21}$$

3.2. The Mixture and the Exponential Connections

After having introduced the Fisher metric and corresponding gradient fields, we now define natural notions of straight lines on M_n and S_{n-1}, respectively, induced by corresponding affine connections. Let us first introduce the straight lines of the so-called mixture connection $\nabla^{(m)}$ on M_n. Given a point $p \in M_n$ and a direction $X \in T_p M_n$, the most natural way to define a straight line that starts in p and has velocity X is given by the so-called m-geodesic

$$\gamma(t) \;=\; p + t\, X \tag{22}$$

We obtain the exponential map for $t = 1$, which is, in this simple example, the translation:

$$\exp_p^{(m)}(X) \;=\; p + X$$

The inverse, therefore, maps a point q to the difference vector that translates p into q:

$$X^{(m)}(p,q) \;:=\; \left(\exp_p^{(m)}\right)^{-1}(q) \;=\; q - p$$

With this difference as X in Equation (22), we obtain the geodesics that connects p with q:

$$\gamma(t) \;=\; p + t\,(q - p) \tag{23}$$

If we choose a point $p \in S_{n-1}$ and $X \in T_p S_{n-1}$, or two points $p, q \in S_{n-1}$, respectively, then the corresponding geodesic Equations (22) and (23) will stay in S_{n-1}. Therefore, the restriction of the exponential map to $T_p S_{n-1}$ and its inverse are trivial:

$$\overline{\exp}_p^{(m)}(X) \;=\; p + X\,, \qquad \overline{X}(p,q) \;:=\; \left(\overline{\exp}_p^{(m)}\right)^{-1}(q) \;=\; q - p$$

where we use a bar over symbols in order to denote the restriction of corresponding objects to S_{n-1}.

Now let us come to the notion of an e-geodesic and the exponential map of the so-called e-connection $\nabla^{(e)}$. Given a point $p \in M_n$ and a direction $X \in T_p M_n$, we consider the geodesic

$$\gamma(t) \;=\; \sum_{i=1}^{n} p_i \exp\left(t\,\frac{X_i}{p_i}\right)\delta_i \tag{24}$$

(The "exp" on the right-hand side of Equation (24) denotes the standard real-valued natural exponential function.) The exponential map of the e-connection is given for $t = 1$:

$$\exp_p^{(e)}(X) \;=\; \sum_{i=1}^{n} p_i \exp\left(\frac{X_i}{p_i}\right)\delta_i$$

with the inverse

$$X^{(e)}(p,q) \;:=\; \left(\exp_p^{(e)}\right)^{-1}(q) \;=\; \sum_{i=1}^{n} p_i \ln\left(\frac{q_i}{p_i}\right)\delta_i$$

This implies that the e-geodesic connecting p with q is given by

$$\gamma(t) \;=\; \sum_{i=1}^{n} p_i \left(\frac{q_i}{p_i}\right)^{t}\delta_i \tag{25}$$

Clearly, if we start in a point $p \in S_{n-1}$ and go along the e-geodesic Equation (24) in a direction X that is tangential to S_{n-1}, we will not stay in S_{n-1}. Analogously, if we connect a point $p \in S_{n-1}$ with a point $q \in S_{n-1}$ in terms of the e-geodesic Equation (25), then the intermediate points will in general not be in the set S_{n-1}. It turns out that, in order to obtain the right exponential map of the e-connection defined on S_{n-1}, we have to normalize the geodesic, which leads to:

$$\overline{\exp}_p^{(e)}(X) \;=\; \sum_{i=1}^{n} \frac{p_i \exp\left(\frac{X_i}{p_i}\right)}{\sum_{j=1}^{n} p_j \exp\left(\frac{X_j}{p_j}\right)}\,\delta_i$$

$$\overline{X}^{(e)}(p,q) \;:=\; \left(\overline{\exp}_p^{(e)}\right)^{-1}(q) \;=\; \sum_{i=1}^{n} p_i \left(\ln\left(\frac{q_i}{p_i}\right) - \sum_{j=1}^{n} p_j \ln\left(\frac{q_j}{p_j}\right)\right)\delta_i$$

3.3. The α-Connections

Given $\alpha \in [-1, 1]$, we define the following convex combination of the mixture connection $\nabla^{(m)}$ and the exponential connection $\nabla^{(e)}$ on M_n:

$$\nabla^{(\alpha)} := \frac{1-\alpha}{2}\nabla^{(m)} + \frac{1+\alpha}{2}\nabla^{(e)} = \nabla^{(m)} + \frac{1+\alpha}{2}\left(\nabla^{(e)} - \nabla^{(m)}\right) \tag{26}$$

The differential equation for the α-geodesic with initial point $p \in M_n$ and initial velocity $X \in T_p M_n$ is given by

$$\ddot{\gamma}_i - \frac{1+\alpha}{2}\frac{\dot{\gamma}_i^2}{\gamma_i} = 0, \qquad \gamma(0) = p, \ \dot{\gamma}(0) = X \tag{27}$$

One can easily verify that Equation (27) is solved by the following curve:

$$\gamma(t) = \sum_{i=1}^{n} p_i \left(1 + t\,\frac{1-\alpha}{2}\,\frac{X_i}{p_i}\right)^{\frac{2}{1-\alpha}} \delta_i \tag{28}$$

By setting $t = 1$, we can define the corresponding α-exponential map:

$$\exp_p^{(\alpha)}(X) = \sum_{i=1}^{n} p_i \left(1 + \frac{1-\alpha}{2}\,\frac{X_i}{p_i}\right)^{\frac{2}{1-\alpha}} \delta_i \tag{29}$$

with the inverse

$$X^{(\alpha)}(p, q) := \left(\exp_p^{(\alpha)}\right)^{-1}(q) = \frac{2}{1-\alpha}\sum_{i=1}^{n} p_i \left(\left(\frac{q_i}{p_i}\right)^{\frac{1-\alpha}{2}} - 1\right)\delta_i \tag{30}$$

Finally, the α-geodesic with initial point p and endpoint q is given by

$$\gamma(t) = \sum_{i=1}^{n} \left(p_i^{\frac{1-\alpha}{2}} + t\left(q_i^{\frac{1-\alpha}{2}} - p_i^{\frac{1-\alpha}{2}}\right)\right)^{\frac{2}{1-\alpha}} \delta_i \tag{31}$$

The α-connection $\overline{\nabla}^{(\alpha)}$ on S_{n-1} is defined as the projection of $\nabla^{(\alpha)}$ with respect to the Fisher metric g. The corresponding geodesic equation is a modification of Equation (27):

$$\ddot{\gamma}_i - \frac{1+\alpha}{2}\left\{\frac{\dot{\gamma}_i^2}{\gamma_i} - \gamma_i \sum_{j=1}^{n}\frac{\dot{\gamma}_j^2}{\gamma_j}\right\} = 0, \qquad \gamma(0) = p, \ \dot{\gamma}(0) = X \tag{32}$$

It is reasonable to make a solution ansatz by normalization of the unconstrained geodesics Equations (28) and (31). However, it turns out that, in order to solve the geodesic Equation (32), both normalized curves have to be reparametrized. More precisely, it has been shown in [8] (Theorems 14.1. and 15.1.) that, with appropriate reparametrizations $\tau_{p,X}$ and $\tau_{p,q}$, we have the following form of the α-geodesic in the simplex S_{n-1}:

$$\gamma_{p,X}(t) = \sum_{i=1}^{n} \frac{p_i \left(1 + \tau_{p,X}(t)\,\frac{1-\alpha}{2}\,\frac{X_i}{p_i}\right)^{\frac{2}{1-\alpha}}}{\sum_{j=1}^{n} p_j \left(1 + \tau_{p,X}(t)\,\frac{1-\alpha}{2}\,\frac{X_j}{p_j}\right)^{\frac{2}{1-\alpha}}}\,\delta_i \tag{33}$$

and

$$\gamma_{p,q}(t) \;=\; \sum_{i=1}^{n} \frac{\left(p_i^{\frac{1-\alpha}{2}} + \tau_{p,q}(t)\left(q_i^{\frac{1-\alpha}{2}} - p_i^{\frac{1-\alpha}{2}}\right)\right)^{\frac{2}{1-\alpha}}}{\sum_{j=1}^{n}\left(p_j^{\frac{1-\alpha}{2}} + \tau_{p,q}(t)\left(q_i^{\frac{1-\alpha}{2}} - p_i^{\frac{1-\alpha}{2}}\right)\right)^{\frac{2}{1-\alpha}}}\,\delta_i \tag{34}$$

Here, the conditions

$$\gamma_{p,X}(0) \;=\; p\,, \quad \dot\gamma_{p,X}(0) \;=\; \dot\tau_{p,X}(0)\,X \;=\; X\,, \qquad \text{and} \qquad \gamma_{p,q}(0) \;=\; p\,, \quad \gamma_{p,q}(1) \;=\; q$$

imply

$$\tau_{p,X}(0) \;=\; 0\,, \quad \dot\tau_{p,X}(0) \;=\; 1\,, \qquad \text{and} \qquad \tau_{p,q}(0) \;=\; 0\,, \quad \tau_{p,q}(1) \;=\; 1$$

Now let us couple X and q by assuming $\gamma_{p,X}(1) \;=\; q$. Together with the condition $\sum_{i=1}^{n} X_i \;=\; 0$, this implies

$$X \;=\; \frac{1}{\tau_{p,X}(1)}\,\frac{2}{1-\alpha}\sum_{i=1}^{n} p_i\left(\frac{\left(\frac{q_i}{p_i}\right)^{\frac{1-\alpha}{2}}}{\sum_{j=1}^{n} p_j\left(\frac{q_j}{p_j}\right)^{\frac{1-\alpha}{2}}} - 1\right)\delta_i \tag{35}$$

Furthermore, if the initial and endpoints of the two curves are identical, then $\gamma_{p,X}(t) = \gamma_{p,q}(t)$ for all t. In particular,

$$\begin{aligned} X \;&=\; \dot\gamma_{p,X}(0) \;=\; \dot\gamma_{p,q}(0) \\ &=\; \dot\tau_{p,q}(0)\,\frac{2}{1-\alpha}\sum_{i=1}^{n} p_i\left(\left(\frac{q_i}{p_i}\right)^{\frac{1-\alpha}{2}} - \sum_{j=1}^{n} p_j\left(\frac{q_j}{p_j}\right)^{\frac{1-\alpha}{2}}\right)\delta_i \end{aligned} \tag{36}$$

A comparison of the Equations (35) and (36) yields

$$\dot\tau_{p,q}(0)\sum_{j=1}^{n} p_j\left(\frac{q_j}{p_j}\right)^{\frac{1-\alpha}{2}} \;=\; \frac{1}{\tau_{p,X}(1)}$$

4. Canonical Divergences for Positive and Probability Measures

4.1. The Relative Entropy (KL-Divergence)

Now we apply the ansatz of Equation (12) in order to define divergence functions for the m- and e-connections on the cone M_n of positive measures. The inverse maps of the corresponding exponential maps are given by

$$\begin{aligned} X^{(m)}(q,p) \;&=\; \sum_{i=1}^{n}(p_i - q_i)\,\delta_i \\ X^{(e)}(q,p) \;&=\; \sum_{i=1}^{n} q_i \ln\frac{p_i}{q_i}\,\delta_i \end{aligned} \tag{37}$$

We can easily verify that the corresponding vector fields

$$q \;\mapsto\; X^{(m)}(q,p)\,, \qquad q \;\mapsto\; X^{(e)}(q,p) \tag{38}$$

are gradient fields: The functions

$$f_i(q) \;:=\; \frac{p_i}{q_i}, \quad \text{and} \quad g_i(q) \;:=\; \ln\frac{p_i}{q_i}$$

trivially satisfy the integrability condition $\frac{\partial f_i}{\partial q_j} = \frac{\partial f_j}{\partial q_i}$ and $\frac{\partial g_i}{\partial q_j} = \frac{\partial g_j}{\partial q_i}$ for all i, j. Therefore, for both connections, there are canonical divergence functions which solve the corresponding Equation (12).

We derive the canonical divergence of the m-connection first, which we denote by $D^{(m)}$. We consider two positive measures p and q and a curve $\gamma{:}[0,1] \to M_n$ connecting q with p, that is $\gamma(0) = q$ and $\gamma(1) = p$. This implies

$$\left\langle X^{(m)}(\gamma(t), p), \dot{\gamma}(t) \right\rangle = \sum_{i=1}^{n} \frac{1}{\gamma_i(t)} \left(p_i - \gamma_i(t)\right) \dot{\gamma}_i(t) \tag{39}$$

and

$$\begin{aligned} D^{(m)}(p \,\|\, q) &= \int_0^1 \left\langle X^{(m)}(\gamma(t), p), \dot{\gamma}(t) \right\rangle dt \\ &= \sum_{i=1}^{n} \int_0^1 \frac{1}{\gamma_i(t)} \left(p_i - \gamma_i(t)\right) \dot{\gamma}_i(t)\, dt \\ &= \sum_{i=1}^{n} \Big[p_i \ln \gamma_i(t) - \gamma_i(t) \Big]_0^1 \\ &= \sum_{i=1}^{n} \left(p_i \ln p_i - p_i \; - \; p_i \ln q_i + q_i \right) \\ &= \sum_{i=1}^{n} \left(q_i - p_i + p_i \ln \frac{p_i}{q_i} \right) \end{aligned}$$

With the same calculation for the e-connection, we obtain the corresponding canonical divergence, which we denote by $D^{(e)}$. Again, we consider a curve γ connecting q with p. This implies

$$\left\langle X^{(e)}(\gamma(t), p), \dot{\gamma}(t) \right\rangle = \sum_{i=1}^{n} \dot{\gamma}_i(t) \, \ln \frac{p_i}{\gamma_i(t)} \tag{40}$$

and

$$\begin{aligned} D^{(e)}(p \,\|\, q) &= \int_0^1 \left\langle X^{(e)}(\gamma(t), p), \dot{\gamma}(t) \right\rangle dt \\ &= \sum_{i=1}^{n} \int_0^1 \dot{\gamma}_i(t) \, \ln \frac{p_i}{\gamma_i(t)}\, dt \\ &= \sum_{i=1}^{n} \left[\gamma_i(t) \left(1 + \ln \frac{p_i}{\gamma_i(t)} \right) \right]_0^1 \\ &= \sum_{i=1}^{n} \left(p_i \; - \; q_i \left(1 + \ln \frac{p_i}{q_i} \right) \right) \\ &= \sum_{i=1}^{n} \left(p_i - q_i + q_i \, \ln \frac{q_i}{p_i} \right) \\ &= D^{(m)}(q \,\|\, p) \end{aligned}$$

These calculations give rise to the following definition:

Definition 1. *The function* $D : M_n \times M_n \to \mathbb{R}$ *defined by*

$$D(p \,\|\, q) := \sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i + \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i} \tag{41}$$

is called the relative entropy *or* Kullback–Leibler divergence. *Its restriction to the set of probability distributions is given by*

$$D(p \,\|\, q) \;:=\; \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i} \tag{42}$$

Proposition 1. *The following holds:*

$$X^{(m)}(q,p) \;=\; -\mathrm{grad}_q\, D(p \,\|\, \cdot)\,, \qquad X^{(e)}(q,p) \;=\; -\mathrm{grad}_q\, D(\cdot \,\|\, p) \tag{43}$$

Furthermore, D is the only function on $M_n \times M_n$ that satisfies the conditions Equation (43) and $D(p \,\|\, p) = 0$ for all p.

Proof. We first compute the partial derivatives

$$\frac{\partial D(p \,\|\, \cdot)}{\partial\, q_i}(q) \;=\; -\frac{p_i}{q_i} + 1\,, \qquad \frac{\partial D(\cdot \,\|\, p)}{\partial\, q_i}(q) \;=\; -\ln \frac{p_i}{q_i}$$

With the Formula (16), we obtain

$$\begin{aligned} \left(\mathrm{grad}_q\, D(p \,\|\, \cdot)\right)_i &\;=\; q_i \left(-\frac{p_i}{q_i} \;+\; 1\right) \;=\; -p_i + q_i \\ \left(\mathrm{grad}_q\, D(\cdot \,\|\, p)\right)_i &\;=\; -q_i \ln \frac{p_i}{q_i} \end{aligned}$$

A comparison with Equation (37) verifies the Equation (43) which uniquely characterize $D(p \,\|\, \cdot)$ as well as $D(\cdot \,\|\, p)$, up to a constant depending on p. With the additional assumption $D(p \,\|\, p) = 0$ for all p, this constant is fixed. □

One can now ask whether the restriction Equation (42) of the Kullback–Leibler divergence to the manifold S_{n-1} is the right divergence function in the sense that Equation (43) also hold for the exponential maps of the restricted m- and e-connections. It is easy to verify that this is indeed the case. Let us elaborate on the geometric reason for this. We consider a general Riemannian manifold M and a submanifold N in it. Given an affine connection ∇ on M, we can define its restriction $\overline{\nabla}$ to N. More precisely, denoting the projection of a vector Z in T_pM onto T_pN by $\Pi_p^\top(Z)$, we define $\overline{\nabla}_X Y\big|_p \;:=\; \Pi_p^\top\left(\nabla_X Y|_p\right)$, where X and Y are vector fields on N. Furthermore, we denote the exponential map of $\overline{\nabla}$ by $\overline{\exp}_p$ and its inverse by $\overline{X}(p,q)$.

Now, given $p \in N$, we consider a function D_p on M, which satisfies the Equation (12). With the restriction $\overline{D}_p$ of D_p to the submanifold N, this directly implies

$$\Pi_q^\top\left(X(q,p)\right) \;=\; -\mathrm{grad}_q\, \overline{D}_p$$

However, in order to have $\overline{X}(q,p) \;=\; -\mathrm{grad}_q\, \overline{D}_p$, which corresponds to the Equation (12) on the submanifold N, the following equality is required:

$$\overline{X}(q,p) \;=\; \Pi_q^\top\left(X(q,p)\right) \tag{44}$$

This condition is satisfied for the m- and e-connections on M_n and its submanifold S_{n-1}, which implies the following proposition.

Proposition 2. *The following holds:*

$$\overline{X}^{(m)}(q,p) \;=\; -\mathrm{grad}_q\, D(p \,\|\, \cdot)\,, \qquad \overline{X}^{(e)}(q,p) \;=\; -\mathrm{grad}_q\, D(\cdot \,\|\, p) \tag{45}$$

where D is given by Equation (42) in Definition 1. Furthermore, D is the only function on $S_{n-1} \times S_{n-1}$ that satisfies the conditions (45) and $D(p \parallel p) = 0$ for all p.

The objects and derivations of this section represent a special case of a general dually flat manifold M, which will be studied in Section 5.

4.2. The α-Divergence

We now extend the method of Section 4.1 to the α-connections, leading to a generalization of the relative entropy, the so-called α-divergence. From the definition of the α-exponential map on the manifold M_n of positive measures, given in Equation (29), we obtain the inverse

$$X^{(\alpha)}(q,p) \;:=\; \left(\exp_q^{(\alpha)}\right)^{-1}(p) \;=\; \frac{2}{1-\alpha}\sum_{i=1}^{n} q_i\left(\left(\frac{p_i}{q_i}\right)^{\frac{1-\alpha}{2}} - 1\right)\delta_i \tag{46}$$

In order to derive the canonical divergence $D^{(\alpha)}$ of the α-connection, which is integrable, we consider two points p and q and a curve $\gamma\colon [0,1] \to M_n$ connecting q with p. We obtain

$$\left\langle X^{(\alpha)}(\gamma(t),p), \dot\gamma(t)\right\rangle \;=\; \frac{2}{1-\alpha}\sum_{i=1}^{n}\dot\gamma_i(t)\left(\left(\frac{p_i}{\gamma_i(t)}\right)^{\frac{1-\alpha}{2}} - 1\right) \tag{47}$$

and

$$\begin{aligned}
D^{(\alpha)}(p \parallel q) &= \int_0^1 \left\langle X^{(\alpha)}(\gamma(t),p), \dot\gamma(t)\right\rangle dt \\
&= \sum_{i=1}^{n}\int_0^1 \frac{2}{1-\alpha}\dot\gamma_i(t)\left(\left(\frac{p_i}{\gamma_i(t)}\right)^{\frac{1-\alpha}{2}} - 1\right) dt \\
&= \sum_{i=1}^{n}\left[\frac{4}{1-\alpha^2}\gamma_i(t)^{\frac{1+\alpha}{2}} p_i^{\frac{1-\alpha}{2}} - \frac{2}{1-\alpha}\gamma_i(t)\right]_0^1 \\
&= \sum_{i=1}^{n}\left(\frac{2}{1+\alpha}p_i - \left(\frac{4}{1-\alpha^2}q_i^{\frac{1+\alpha}{2}} p_i^{\frac{1-\alpha}{2}} - \frac{2}{1-\alpha}q_i\right)\right) \\
&= \sum_{i=1}^{n}\left(\frac{2}{1-\alpha}q_i + \frac{2}{1+\alpha}p_i - \frac{4}{1-\alpha^2}q_i^{\frac{1+\alpha}{2}} p_i^{\frac{1-\alpha}{2}}\right)
\end{aligned}$$

Obviously, we have

$$D^{(-\alpha)}(p \parallel q) \;=\; D^{(\alpha)}(q \parallel p) \tag{48}$$

These calculations give rise to the following definition:

Definition 2. *The function $D^{(\alpha)} : M_n \times M_n \to \mathbb{R}$ defined by*

$$D^{(\alpha)}(p \parallel q) \;:=\; \frac{2}{1-\alpha}\sum_{i=1}^{n} q_i + \frac{2}{1+\alpha}\sum_{i=1}^{n} p_i - \frac{4}{1-\alpha^2}\sum_{i=1}^{n} q_i^{\frac{1+\alpha}{2}} p_i^{\frac{1-\alpha}{2}} \tag{49}$$

*is called the α-*divergence*. Its restriction to probability measures is given as*

$$D^{(\alpha)}(p \parallel q) \;=\; \frac{4}{1-\alpha^2}\left(1 - \sum_{i=1}^{n} q_i^{\frac{1+\alpha}{2}} p_i^{\frac{1-\alpha}{2}}\right)$$

Proposition 3. *The following holds:*

$$X^{(\alpha)}(q,p) \;=\; -\mathrm{grad}_q\, D^{(\alpha)}(p \parallel \cdot) \tag{50}$$

Furthermore, $D^{(\alpha)}$ is the only function on $M_n \times M_n$ that satisfies the condition (50) and $D^{(\alpha)}(p \,\|\, p) = 0$ for all p.

Proof. We compute the partial derivative

$$\frac{\partial D^{(\alpha)}(p \,\|\, \cdot)}{\partial q_i}(q) \;=\; \frac{2}{1-\alpha}\left(1 - q_i^{\frac{1+\alpha}{2}-1} p_i^{\frac{1-\alpha}{2}}\right)$$

With the Formula (16), we obtain

$$\begin{aligned}\left(\operatorname{grad}_q D^{(\alpha)}(p \,\|\, \cdot)\right)_i &= q_i \cdot \frac{2}{1-\alpha}\left(1 - q_i^{\frac{1+\alpha}{2}-1} p_i^{\frac{1-\alpha}{2}}\right)\\ &= \frac{2}{1-\alpha}\left(q_i - q_i^{\frac{1+\alpha}{2}} p_i^{\frac{1-\alpha}{2}}\right)\end{aligned}$$

A comparison with Equation (46) verifies Equation (50) which uniquely characterizes $D^{(\alpha)}(p \,\|\, \cdot)$, up to a constant depending on p. With the additional assumption $D^{(\alpha)}(p \,\|\, p) = 0$ for all p, this constant is fixed. □

In what follows, we use the notation $D^{(\alpha)}$ also for $\alpha \in \{-1, 1\}$ by setting $D^{(-1)}(p \,\|\, q) := D(p \,\|\, q)$ and $D^{(1)}(p \,\|\, q) := D(q \,\|\, p)$ where D is relative entropy defined by Equation (41). This is consistent with the definition of the α-connections, given by Equation (26), where we have the m-connection for $\alpha = -1$ and the e-connection for $\alpha = 1$. Note that $D^{(0)}$ is closely related to the Hellinger distance

$$d^H(p, q) \;:=\; \left(\sum_{i=1}^{n}\left(p_i^{\frac{1}{2}} - q_i^{\frac{1}{2}}\right)^2\right)^{\frac{1}{2}}$$

More precisely, we have

$$D^{(0)}(p \,\|\, q) \;=\; 2\left(d^H(p, q)\right)^2 \tag{51}$$

In fact, the derivation of $D^{(\alpha)}$ was based on the idea to associate a distance-like function to the α-connections through the general Equation (12). However, it turns out that, although being naturally motivated, the functions $D^{(\alpha)}$ do not share all properties of the square of a distance, except for $\alpha = 0$. The symmetry is obviously not satisfied. On the other hand, we have $D^{(\alpha)}(p \,\|\, q) \geq 0$, and $D^{(\alpha)}(p \,\|\, q) = 0$ if and only if $p = q$.

We now ask whether the restriction of $D^{(\alpha)}$, which is defined for positive measures, to the simplex S_{n-1} of probability distributions is the canonical divergence for the α-connections on S_{n-1}. We have seen that this is the case for the m- and e-connections, that is for $\alpha \in \{-1, +1\}$. However, for general α, the situation is more complicated. From Equation (36) we obtain

$$\overline{X}^{(\alpha)}(q, p) \;=\; \dot{\tau}_{q,p}(0)\, \Pi_q^{\top}\left(X^{(\alpha)}(q, p)\right)$$

This equality deviates from the condition of Equation (44) by the factor $\dot{\tau}_{q,p}(0)$, which proves that the restriction of the α-divergence to S_{n-1} does not coincide with the canonical α-divergence on the simplex. As an example, we consider the case $\alpha = 0$, where the α-connection is the Levi-Civita connection of the Fisher metric. As we will see in the next section, the canonical divergence in that case equals $\overline{D}^{(0)}(p \,\|\, q) \;=\; \frac{1}{2}\left(d^F(p, q)\right)^2$, where d^F denotes the distance with respect to the Fisher metric (see Equation (62)). Obviously, this divergence is different from the divergence $D^{(0)}$, given by Equation (51), which is based on the distance in the ambient space M_n, the Hellinger distance.

5. General Canonical Divergence and Its Consistency

5.1. Canonical Divergence

We have derived a canonical divergence when the vector field X of the inverse exponential map, that is $\exp_q(X(q,p)) = p$ for all p and q, is integrable. We now define a canonical divergence in a general n-dimensional dual manifold (M, g, ∇, ∇^*). Consider a ∇-geodesic $\gamma_{q,p}{:}[0,1] \to M$ connecting q and p. We define a tangent vector field $X_t(p,q)$ along this geodesic:

$$X_t(q,p) := X\left(\gamma_{q,p}(t), p\right) \tag{52}$$

Obviously,

$$X_0 = X(q,p) \tag{53}$$

$$X_1(q,p) = 0 \tag{54}$$

Definition 3. *A canonical divergence from p to q is defined by the path integral*

$$D(p \,\|\, q) = \int_0^1 \left\langle X_t(q,p), \dot{\gamma}_{q,p}(t) \right\rangle \, dt \tag{55}$$

Replacing the ∇-geodesic $\gamma_{q,p}$ from q to p by the reversed ∇-geodesic $\gamma_{p,q}$ from p to q and the vector field $X_t(q,p)$ by the vector field $X_t^*(p,q) := X^*\left(\gamma_{p,q}(t), p\right)$ of the dual connection ∇^* leads to the following related definition of a canonical divergence:

$$D'(p \,\|\, q) := \int_0^1 \left\langle X_t^*(p,q), \dot{\gamma}_{p,q}(t) \right\rangle \, dt \tag{56}$$

$$= -\int_0^1 \left\langle X^*(\gamma_{q,p}(t), q), \dot{\gamma}_{q,p}(t) \right\rangle \, dt \tag{57}$$

Although motivated and derived in different terms, the divergence of the article [9] turns out to coincide with D'. The authors apply Hooke's law to a "∇^*-spring" and define their divergence, in terms of an expression related to Equation (57), as the work that is necessary to move a point of unit mass from q to p along the ∇-geodesic $\gamma_{q,p}$ against the force field $X^*(\gamma_{q,p}(t), q)$. We became aware of this article after submission of our present article. The divergence D' shares many nice properties of our canonical divergence. However, in the integrability case, it is not generally true that $X(q,p) = -\mathrm{grad}_q D'(p \,\|\, \cdot)$, a property that serves as main motivation of our article and which is satisfied by our canonical divergence of Equation (55).

Before stating the main result that the canonical divergence defined by Equation (55) induces the same Riemannian metric g and the same pair of affine connections ∇ and ∇^*, we show some of its properties. Since the geodesic connecting $\gamma_{q,p}(t)$ and p is a part of the geodesic connecting q and p, corresponding to the interval $[t,1]$, the inverse exponential map at $\gamma_{q,p}(t)$ satisfies

$$X_t\left(q,p\right) = (1-t)\, \dot{\gamma}_{q,p}(t) \tag{58}$$

Hence, we have

$$D(p \,\|\, q) = \int_0^1 (1-t) \, \left\| \dot{\gamma}_{q,p}(t) \right\|^2 \, dt \tag{59}$$

where

$$\left\| \dot{\gamma}_{q,p}(t) \right\|^2 = \left\langle \dot{\gamma}_{q,p}(t), \dot{\gamma}_{q,p}(t) \right\rangle \tag{60}$$

This already proves $D(p \,\|\, q) \geq 0$, and $D(p \,\|\, q) = 0$ if and only if $p = q$. If we replace the parameter t by $1-t$ and use $\gamma_{q,p}(t) = \gamma_{p,q}(1-t)$, we directly obtain the following representation of the canonical divergence:

Proposition 4. *The divergence of Definition 3 is given by*

$$D(p \,\|\, q) \;=\; \int_0^1 t \,\left\|\dot{\gamma}_{p,q}(t)\right\|^2 \,dt \tag{61}$$

where $\gamma_{p,q}$ denotes the geodesic from p to q.

Remark 1. *In the special case where M is self-dual, $\nabla = \nabla^*$ is the Levi-Civita connection with respect to g. In that case, the velocity field $\dot{\gamma}_{p,q}$ is parallel along the geodesic $\gamma_{p,q}$, and therefore*

$$\left\|\dot{\gamma}_{p,q}(t)\right\|_{\gamma(t)} \;=\; \left\|\dot{\gamma}_{p,q}(0)\right\|_p \;=\; \left\|X(p,q)\right\|_p \;=\; d(p,q)$$

where $d(p,q)$ denotes the Riemannian distance between p and q. This implies that the canonical divergence corresponds to the energy of the geodesic $\gamma_{p,q}$, that is

$$D(p \,\|\, q) \;=\; \frac{1}{2}\, d^2(p,q) \tag{62}$$

In the general case, where ∇ is not necessarily the Levi-Civita connection, we obtain the energy of the geodesic $\gamma_{p,q}$ as the symmetrized version of the canonical divergence:

$$\frac{1}{2}\big(D(p \,\|\, q) + D(q \,\|\, p)\big) \;=\; \frac{1}{2}\int_0^1 \left\|\dot{\gamma}_{p,q}(t)\right\|^2 \,dt \tag{63}$$

Remark 2. *Let us compare the canonical divergence D of the affine connection ∇ with the canonical divergence D^* of its dual connection ∇^*, both defined by Equation (55) or equivalently by Equation (61). In the special case of the α-connection $\nabla = \nabla^{(\alpha)}$, we have $D^*(p \,\|\, q) = D(q \,\|\, p)$ (see Equation (48)). In Section 5.3, we will prove that this kind of symmetry holds in the general case of a dually flat manifold. However, our canonical divergence does not necessarily have this property, when the space is not dually flat. This is contrary to most other approaches where the symmetry is considered to be a natural property of any divergence. In order to have that property also in our setting, we can consider the mean canonical divergence*

$$D^{\nabla}_{mcd}(p \,\|\, q) \;:=\; \frac{1}{2}\big(D(p \,\|\, q) + D^*(q \,\|\, p)\big) \tag{64}$$

which obviously satisfies

$$D^{(\nabla^*)}_{mcd}(p \,\|\, q) \;=\; D^{\nabla}_{mcd}(q \,\|\, p) \tag{65}$$

As we will prove in the next section, the canonical divergence D induces the metric g and the connections ∇ and ∇^*. The same holds for the mean canonical divergence D^{∇}_{mcd}. However, if ∇ is integrable, then it is not generally true that $X(q,p) = -\mathrm{grad}_q\, D^{\nabla}_{mcd}(p \,\|\, \cdot)$, which is inconsistent with the main motivation of our canonical divergence (see Equation (12)).

5.2. Main Consistency Result

Let $\overset{D}{g}$, $\overset{D}{\nabla}$, and $\overset{D}{\nabla}{}^*$ be the geometrical objects derived from the canonical divergence D as defined in Equation (55). We recall the corresponding definitions from Section 1 in terms of a local coordinate system $\boldsymbol{\xi} = (\xi^1, \cdots, \xi^n)$:

$$\overset{D}{g}_{ij}(p) \;=\; \partial'_i \partial'_j\, D(\boldsymbol{\xi}_p \,\|\, \boldsymbol{\xi}_q)\Big|_{q=p} \tag{66}$$

$$\overset{D}{\Gamma}_{ijk}(p) \;=\; -\partial_i \partial_j \partial'_k\, D\left(\boldsymbol{\xi}_p \,\|\, \boldsymbol{\xi}_q\right)\Big|_{q=p} \tag{67}$$

$$\overset{D}{\Gamma}{}^*_{ijk}(p) \;=\; -\partial'_i \partial'_j \partial_k\, D\left(\boldsymbol{\xi}_p \,\|\, \boldsymbol{\xi}_q\right)\Big|_{q=p} \tag{68}$$

We have defined our canonical divergence D based on a metric g and an affine connection ∇. It is natural to require that this divergence is consistent in the sense that the objects $\overset{D}{g}$, $\overset{D}{\nabla}$, and $\overset{D}{\nabla}{}^*$ coincide with the original objects g, ∇, and ∇^* of M, where ∇^* is the dual affine connection of ∇ with respect to g. Since the geometry is determined by the derivatives of $D\left(\boldsymbol{\xi}_p \,\|\, \boldsymbol{\xi}_q\right)$ at $p = q$, we consider the case where p and q are close to each other, that is

$$z^i = \xi^i_q - \xi^i_p \tag{69}$$

is small for all i. We evaluate the divergence by Taylor expansion up to $O\left(\|\boldsymbol{z}\|^3\right)$. Note that $X(p,q)$ is of order $\|\boldsymbol{z}\|$.

Proposition 5. *When $\|\boldsymbol{z}\| = \|\boldsymbol{\xi}_q - \boldsymbol{\xi}_p\|$ is small, the canonical divergence is expanded as*

$$D(p \,\|\, q) \;=\; \frac{1}{2}\, g_{ij}(p)\, z^i z^j + \frac{1}{6}\, \Lambda_{ijk}(p)\, z^i z^j z^k + O\left(\|\boldsymbol{z}\|^4\right) \tag{70}$$

where

$$\Lambda_{ijk} \;=\; 2\,\partial_i g_{jk} - \Gamma_{ijk} \tag{71}$$

Proof. We obtain the local coordinates $\boldsymbol{\xi}(t)$ of the geodesic $\gamma_{p,q}(t)$ in Taylor series as

$$\xi^i(t) = \xi^i_p + t\, X^i - \frac{t^2}{2}\, \Gamma^i_{jk} X^j X^k + O\left(\|tX\|^3\right) \tag{72}$$

where $X^i = X^i(p,q)$. When $\boldsymbol{z}$ is small, X is of order $O(\|\boldsymbol{z}\|)$. Hence, we regard Equation (72) as Taylor expansion with respect to X, and $t \in [0,1]$ when $\boldsymbol{z}$ is small. When $t = 1$, we have

$$z^i = X^i - \frac{1}{2}\, \Gamma^i_{jk} X^j X^k \tag{73}$$

where the higher-order terms are neglected. This in turn gives

$$X^i = z^i + \frac{1}{2}\, \Gamma^i_{jk} z^j z^k \tag{74}$$

We calculate $D(p \,\|\, q)$ by using Equation (61). The velocity at t is given as

$$\dot{\xi}^i(t) \;=\; X^i - t\, \Gamma^i_{jk} X^j X^k \tag{75}$$

$$\;=\; z^i + \frac{1}{2}(1 - 2t)\, \Gamma^i_{jk} z^j z^k \tag{76}$$

We also use

$$g_{ij}\left(\boldsymbol{\xi}(t)\right) = g_{ij}\left(\boldsymbol{\xi}_p\right) + t\, \partial_k g_{ij} z^k \tag{77}$$

Collecting these terms, we have

$$t\, g_{ij}\left(\boldsymbol{\xi}(t)\right)\, \dot{\xi}^i(t)\, \dot{\xi}^j(t) \;=\; t\, g_{ij}\, z^i z^j + \left\{ t^2\, \partial_i g_{jk} + \left(-2t^2 + t\right) \Gamma_{ijk} \right\}\, z^i z^j z^k \tag{78}$$

By integration, we have

$$D(p \,\|\, q) \;=\; \int_0^1 t\, g_{ij}\left(\boldsymbol{\xi}(t)\right)\, \dot{\xi}^i(t)\, \dot{\xi}^j(t)\, dt \tag{79}$$

$$\;=\; \frac{1}{2}\, g_{ij} z^i z^j + \frac{1}{6}\, \Lambda_{ijk}\, z^i z^j z^k \tag{80}$$

where indices of Λ_{ijk} are symmetrized because of multiplication of $z^i z^j z^k$. This gives Equation (70). □

Theorem 1. (Consistency theorem) *The geometric quantities $\overset{D}{g}$, $\overset{D}{\nabla}$, and $\overset{D}{\nabla}{}^*$, derived from the canonical divergence $D(p \parallel q)$ of Definition 3 coincide with the original quantities g, ∇, and ∇^*.*

Proof. By differentiating Equation (70) with respect to $\boldsymbol{\xi}_p$,

$$\partial_i D = \frac{1}{2}\partial_i g_{jk}\, z^j z^k - g_{ij}\, z^j - \frac{1}{2}\,\Lambda_{ijk}\, z^j z^k \tag{81}$$

$$\partial_i \partial_j D = \frac{1}{2}\,\partial_i \partial_j g_{kl}\, z^k z^j - 2\,\partial_i g_{jk}\, z^k + g_{ij} + \Lambda_{ijk}\, z^k \tag{82}$$

of which the indexed quantities of the right-hand side need to be symmetrized with respect to i, j. By evaluating $\partial_i \partial_j D$ at $\boldsymbol{\xi}_p = \boldsymbol{\xi}_q$, i.e., $\boldsymbol{z} = 0$, we have

$$\overset{D}{g}_{ij} = g_{ij} \tag{83}$$

proving that the Riemannian metric derived from D is the same as the original one. We further differentiate Equation (82) with respect to $\boldsymbol{\xi}_q$ and evaluate it at $\boldsymbol{\xi}_p = \boldsymbol{\xi}_q$. This yields

$$\overset{D}{\Gamma}_{ijk} = -\partial_i \partial_j \partial'_k D = 2\,\partial_i g_{jk} - \Lambda_{ijk} \tag{84}$$

$$= \Gamma_{ijk} \tag{85}$$

Hence, the affine connection $\overset{D}{\nabla}$ derived from D is exactly the same as the original affine connection ∇. □

Remark 3. *In the special case $\nabla = \nabla^*$, the canonical divergence is given by half of the squared norm of the inverse exponential map (see Equation (62)):*

$$D(p \parallel q) = \frac{1}{2}\,\|X(p,q)\|_p^2 \tag{86}$$

The right-hand side of Equation (86) defines a divergence for a general connection, which coincides with the canonical divergence in the self-dual case. We have studied this divergence in our previous work [6]. We have shown that this divergence recovers g in terms of Equation (66). However, it fails to recover ∇ and ∇^* in terms of Equations (67) and (68) directly. In order to overcome this shortcoming, we considered the α-connection $\nabla^{(\alpha)} = \frac{1-\alpha}{2}\,\nabla + \frac{1+\alpha}{2}\,\nabla^*$ and the corresponding inverse exponential map $X^{(\alpha)}$, which imply the following version of Equation (86):

$$D^{(\alpha)}(p \parallel q) := \frac{1}{2}\,\|X^{(\alpha)}(p,q)\|_p^2 \tag{87}$$

($D^{(\alpha)}$ does not denote the α-divergence here.) We have shown in [6] that for $\alpha = -\frac{1}{3}$ the divergence $D^{(\alpha)}$, referred to it as standard divergence, *induces the original quantities g, ∇, and ∇^*. It turns out, however, that this first attempt to define a canonical divergence has serious limitations. For instance, it does not reduce to the known canonical divergence in the dually flat case. This important property is satisfied by the canonical divergence of Definition 3, which we are going to prove in the next section.*

5.3. Canonical Divergence in a Dually Flat Manifold

When a manifold M is dually flat, it has an affine coordinate system $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^n)$ and a potential function $\psi(\boldsymbol{\theta})$, where the dual affine coordinates $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ are given by

$$\eta_i = \frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta^i}, \qquad i = 1, \ldots, n \tag{88}$$

The dual potential is then defined as

$$\varphi(\boldsymbol{\eta}) = \psi(\boldsymbol{\theta}) - \boldsymbol{\theta} \cdot \boldsymbol{\eta} \tag{89}$$

where $\boldsymbol{\theta} \cdot \boldsymbol{\eta} = \theta^i \eta_i$ and $\boldsymbol{\theta}$ is a function of $\boldsymbol{\eta}$ by Equation (88). The geodesic connecting p and q, a generalisation of the e-geodesic of Section 3.2, has the form

$$\boldsymbol{\theta}(t) = \boldsymbol{\theta}_p + t\left(\boldsymbol{\theta}_q - \boldsymbol{\theta}_p\right) \tag{90}$$

Hence, the velocity is constant

$$\dot{\boldsymbol{\theta}}(t) = \boldsymbol{z} = \boldsymbol{\theta}_q - \boldsymbol{\theta}_p \tag{91}$$

The canonical divergence from $\boldsymbol{\theta}_p$ to $\boldsymbol{\theta}_q$ is defined by

$$D\left(\boldsymbol{\theta}_p \,\|\, \boldsymbol{\theta}_q\right) = \int_0^1 t\, g_{ij}\left(\boldsymbol{\theta}(t)\right) z^i z^j dt \tag{92}$$

Since $g_{ij} = \partial_i \partial_j \psi$, we have

$$\begin{aligned} D\left(\boldsymbol{\theta}_p \,\|\, \boldsymbol{\theta}_q\right) &= \int_0^1 t\, \partial_i \partial_j \psi\left(\boldsymbol{\theta}_p + t\,\boldsymbol{z}\right) z^i z^j dt && (93)\\ &= \int_0^1 t\, \ddot{\psi}\left(\boldsymbol{\theta}(t)\right) dt && (94)\\ &= -\int_0^1 \dot{\psi}\left(\boldsymbol{\theta}(t)\right) dt + \left[t\, \dot{\psi}\left(\boldsymbol{\theta}(t)\right)\right]_0^1 && (95)\\ &= \psi\left(\boldsymbol{\theta}_p\right) + \varphi\left(\boldsymbol{\eta}_q\right) - \boldsymbol{\theta}_p \cdot \boldsymbol{\eta}_q && (96) \end{aligned}$$

This shows that our canonical divergence is the same as the canonical divergence defined in terms of the Bregman divergence of M.

Now we come back to the symmetry property that we already addressed in Remark 2. We derived $D(p \,\|\, q)$ by using the primal affine connection ∇ and the related inverse exponential map. We can construct its dual $D^*(p \,\|\, q)$ by using the dual affine connection ∇^* and the dual inverse exponential map. The dual affine coordinates are $\boldsymbol{\eta}$, and the m-geodesic connecting p and q is given by

$$\boldsymbol{\eta}(t) = \boldsymbol{\eta}_p + t\left(\boldsymbol{\eta}_q - \boldsymbol{\eta}_p\right) \tag{97}$$

Hence, the velocity is constant

$$\dot{\boldsymbol{\eta}}(t) = \boldsymbol{z}^* = \boldsymbol{\eta}_q - \boldsymbol{\eta}_p \tag{98}$$

The dual canonical divergence D^* is defined by

$$D^*(p \,\|\, q) = \int_0^1 t\, g^{ij}\left(\boldsymbol{\eta}_t\right) z_i^* z_j^*\, dt \tag{99}$$

Here,

$$g^{ij}(\boldsymbol{\eta}) = \partial^i \partial^j \varphi(\boldsymbol{\eta}) \tag{100}$$

where

$$\partial^i = \frac{\partial}{\partial \eta_i} \tag{101}$$

So we have

$$D^*(p \,\|\, q) = \int_0^1 t\, \partial^i \partial^j \varphi \left(\boldsymbol{\eta}_p + t\boldsymbol{z}^*\right) z_i^* z_j^*\, dt \tag{102}$$

By similar calculations, we have

$$D^*(p \,\|\, q) \;=\; D(q \,\|\, p) \tag{103}$$

This proves that ∇ and ∇^* give the same canonical divergence except that p and q are interchanged because of the duality. Such a nice property holds when M is dually flat.

6. Geodesic Projections and Integrability

Given a divergence D on M and a point $p \in M$, we consider the set of points q that satisfy

$$D(p \,\|\, q) \;=\; \text{const} \tag{104}$$

where p is fixed. This set is the surface of the equi-divergence ball centered at p. When a smooth submanifold S is given, we search for a point $\hat{p} \in S$ that minimizes $D(p \,\|\, q)$, $q \in S$. Intuitively, we obtain such a minimizer by considering a ball centered at p. We increase its radius, starting from 0, until the ball touches S for the first time. Any touch point $\hat{p}$ is then a minimizer of $D(p \,\|\, q)$, $q \in S$. When the geodesic connecting $\hat{p}$ and p is orthogonal to S at $\hat{p}$, we call $\hat{p}$ a *geodesic projection* of p onto S.

Definition 4. *We say that the* geodesic projection property *holds if every minimizer $\hat{p}$ of the divergence D is given by the geodesic projection of p onto S.*

We know that the geodesic projection property holds when M is dually flat, but it does not hold in general. The following condition guarantees the geodesic projection property:

Proposition 6. *The geodesic projection property holds when the inverse exponential map $X(q, p)$ is in proportion to the gradient of $D(p \,\|\, q)$ with respect to q,*

$$X(q, p) \;=\; c \cdot \operatorname{grad}_q D(p \,\|\, \cdot) \tag{105}$$

where c is a constant that may depend on q and p.

Proof. Consider the geodesic connecting $q = \hat{p}$ and p. Then, the tangent vector at q is $X(q, p)$. Assume that $X(q, p)$ has the same direction as the gradient $\operatorname{grad}_q D(p \,\|\, \cdot)$, that is, the vector orthogonal to the surface of the ball touching S. Then $X(q, p)$ is also orthogonal to the tangent space of S in $\hat{p}$, as the tangent space of the ball contains the tangent space of S at this point. This means that $\hat{p}$ is a geodesic projection. □

Obviously, when the vector field of the inverse exponential map is integrable, the geodesic projection property directly follows from Equation (12). We have shown that this intergrability condition is satisfied for general dually flat manifolds. In particular, the integrability is satisfied for the α-connection $\nabla^{(\alpha)}$ defined on the cone M_n of positive measures, which leads to the α-divergence as canonical divergence. The restriction of the α-connection to the simplex S_{n-1} of probability distributions, denoted by $\overline{\nabla}^{(\alpha)}$, is still integrable, even though S_{n-1} is not (dually) flat with respect $\overline{\nabla}^{(\alpha)}$ if $\alpha \notin \{-1, +1\}$. As we have seen, the canonical divergence associated with $\overline{\nabla}^{(\alpha)}$ does not coincide with the restriction of the α-divergence to S_{n-1}. However, this restriction is still useful in the context of applications that require projections onto submanifolds S. The reason is that the geodesic

projection property holds for $\overline{\nabla}^{(\alpha)}$. To be more precise, consider the restriction of the α-divergence to the simplex S_{n-1}:

$$D^{(\alpha)}(p \,\|\, q) \;=\; \frac{4}{1-\alpha^2}\left(1 - \sum_{i=1}^{n} q_i^{\frac{1+\alpha}{2}} p_i^{\frac{1-\alpha}{2}}\right)$$

The gradient is given as

$$\operatorname{grad}_q D^{(\alpha)}(p \,\|\, \cdot) \;=\; -\frac{2}{1-\alpha} \sum_i q_i \left(\left(\frac{p_i}{q_i}\right)^{\frac{1-\alpha}{2}} - \sum_j q_j \left(\frac{p_j}{q_j}\right)^{\frac{1-\alpha}{2}}\right) \delta_i$$

Comparing this with Equation (36) we see that

$$X(q,p) \;=\; -\dot{\tau}_{q,p}(0) \operatorname{grad}_q D^{(\alpha)}(p \,\|\, \cdot)$$

With the condition (105) this implies that the geodesic projection property holds for $D^{(\alpha)}$, even though it is not the canonical α-divergence on the simplex.

Author Contributions: The research was designed and carried out by both authors. They both wrote the paper, with main contribution by Nihat Ay. Both authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Eguchi, S. Second order efficiency of minimum contrast estimators in a curved exponential family. *Ann. Stat.* **1983**, *11*, 793–803.
2. Amari, S.-I.; Nagaoka, H. *Methods of Information Geometry*; American Mathematical Society: Providence, RI, USA; Oxford University Press: Oxford, UK, 2000.
3. Matumoto, T. Any statistical manifold has a contrast function—On the C^3-functions taking the minimum at the diagonal of the product manifold. *Hiroshima Math. J.* **1993**, *23*, 327–332.
4. Kurose, T. On the divergence of 1-conformally flat statistical manifolds. *Tohoku Math. J.* **1994**, *46*, 427–433.
5. Matsuzoe, H. On realization of conformally-projectively flat statistical manifolds and the divergences. *Hokkaido Math. J.* **1998**, *27*, 409–421.
6. Amari, S.-I.; Ay, N. Standard Divergence in Manifold of Dual Affine Connections. In *Geometric Science of Information*, Proceedings of the 2nd International Conference on Geometric Science of Information, Palaiseau, France, 28–30 October 2015.
7. Hofbauer, J.; Sigmund, K. *Evolutionary Games and Population Dynamics*; Cambridge University Press: Cambridge, UK, 2002.
8. Morozova, E.A.; Chentsov, N.N. Markov invariant geometry on manifolds of states. *J. Sov. Math.* **1991**, *56*, 2648–2669.
9. Henmi, M.; Kobayashi, R. Hooke's law in statistical manifolds and divergences. *Nagoya Math. J.* **2000**, *159*, 1–24.